DOCUMENT RESUME

ED 227 495 CS 207 397

AUTHOR Lees, Elaine O.
TITLE Critical Monism, Critical Pluralism, and the Ideal of Inter-Rater Reliability.
PUB DATE Mar 81
NOTE 13p.; Paper presented at the Annual Meeting of the Conference on College Composition and Communication (32nd, Dallas, TX, March 26-28, 1981).
PUB TYPE Viewpoints (120) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Elementary Secondary Education; *Evaluation Criteria; Evaluation Methods; *Evaluative Thinking; Evaluators; *Interrater Reliability; Reliability; Theories; *Writing Evaluation
IDENTIFIERS *Critical Theory

ABSTRACT

Given the concern for reliability in essay evaluation and the prospect of "error" variance in its absence, methods to promote interrater reliability in the evaluation of written compositions have been developed. These methods reduce variation in the value systems being applied by readers to texts, either by limiting the group of readers to those who are likely to share value systems to begin with or by encouraging readers toward conformity by making them state explicitly the qualities in texts they will reward with high scores or penalize with low ones. Two positions within critical theory that are relevant to the issue of rater reliability in essay scoring are critical monism (asserting that a text can be said to have only one correct meaning) and critical pluralism (asserting that a text can have more than one correct meaning). According to critical monism, if readers could identify an author's intentions reliably and use relative readability to measure how successfully the intentions were realized through the text, these indexes could be used as a means of certifying readers to evaluate student writing. According to the critical pluralism theory, however, one's sense of an author's intention is always traceable to one's interpretive strategy. And since more than one legitimate interpretive strategy is open to readers, so is more than one sense of an author's intention. Thus, texts can be stabilized by stabilizing readers and by teaching them a common interpretive strategy. The important issue is whether the group's judgments can be taken to represent the judgments of the set of competent readers as a whole. Unless these interpretations are identified, the variance arising from differences in readers's interpretive communities cannot be distinguished from random variance. (HOD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED227495

CCCC 1981
Session I-7
"Achieving Reliability
in Holistic Evaluation"

CRITICAL MONISM, CRITICAL PLURALISM, AND THE IDEAL OF INTER-RATER RELIABILITY

Elaine O. Lees
University of Pittsburgh

For more than a century--that is, for the majority of the time English studies has existed as a formal discipline in universities--teachers in this field have possessed evidence that, left to their own devices, they graded essay tests unreliably. In her review of the literature on essay rating, Edith Huddleston in 1954 noted that studies as far back as the late nineteenth century showed "considerable discrepancy" among different teachers' evaluations of the same essay and that "a vast array of later studies confirm[ed] this fact" (Huddleston, p. 165). Teachers not only varied among themselves in the ratings they assigned to a single essay; when circumstances permitted, they also varied as individuals, assigning the same essay a high mark on one occasion and a low mark on another (Stalnaker and Stalnaker, p. 603). In a study of the evaluation of CEEB essay questions in 1936, John Stalnaker reported that if, in lieu of rating by a second reader, each essay in the batch were arbitrarily assigned a 'D' grade, the level of agreement between first and second ratings would drop to 37 percent--a level only slightly lower than the 41 percent level actually achieved. And in a well known and more recent study of the evaluation patterns of educated readers in business, law, editing, and the sciences--as well as English--Paul Diederich and his associates found that a third of the 300 papers received every grade their nine-level scale allowed; 94 percent of the papers received at least seven different grades; and no paper received less than five different grades. Though the English teachers who read for Diederich agreed among themselves slightly more frequently than did the members of other professional groups, the degree of correlation they achieved was still an unimpressive .41 (McColly, p. 150).

CS207397

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Elaine O. Lees

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

In what can be read as an ironic footnote to such studies, Huddleston reports that "a review of the literature indicates . . . the unreliability of essay examinations is most pronounced in the area of English composition," while attempts to rate reliably essays in "content" fields like history or physics "have been somewhat more successful" (p. 165). The scoring of writing seems most reliable, that is, when it's not writing that's being scored.

Almost universally, members of the profession have greeted low levels of inter-rater reliability like these with dismay. Stalnaker (1937) comments: "Such discrepancies make it impossible to get an index of the candidate's ability to write" (p. 135), and states elsewhere: "No test of low reliability can possibly be valid" (S & S, 599). More recently, in The Philosophy of Composition (1977), E. D. Hirsch, Jr., names the development of reliable methods of evaluation as "the sine qua non of all future composition research, since without it we have no way of testing anything else" (p. 172). The prevailing view of inter-rater reliability may be summed up in the words of Julian Stanley in the American Council on Education's volume, Educational Measurement (1971, 2nd ed). Stanley writes:

> In essay examinations . . . or in fact anything that calls for interpretation or evaluation by an observer or scorer, variance due to the scorer will enter in. There will be variance associated with different scorers, variance associated with changes in each scorer from time to time, and variance representing unpredictable within-scorer inconsistency. These types of variance . . . become important in certain particular measurement situations; where they occur they ordinarily constitute error variance, and procedures should be planned to identify them and treat them as error. (p. 368)

It is not surprising--given a century of concern for reliability in essay evaluation and the prospect of "error" variance in its absence--that methods to promote inter-rater reliability in the evaluation of written compositions have been developed. Such methods ordinarily include some combination of the following seven practices:

1. Raters with a common background are chosen.

2. Raters pass a qualifying test in order to be eligible to read.
3. Raters agree to share a set of common evaluative criteria, either criteria they have formulated themselves or criteria an investigator supplies.
4. Raters study a common set of scored sample essays before beginning to evaluate on their own.
5. Raters practice reading unscored sample essays together.
6. Raters discuss differences that arise in their evaluations of sample essays.
7. Raters are monitored during the actual evaluation session to ensure that they're reading consistently.

Investigators who have used such methods generally report levels of inter-rater reliability much higher than the figures I cited earlier. The success of their methods seems to promise alleviation of the "reliability problem." And in fact, Charles Cooper affirms that, suitably chosen and trained, raters now "can achieve nearly perfect agreement in choosing the better of a pair of essays; and they can achieve scoring reliabilities in the high eighties and low nineties on their summed scores from multiple pieces of a student's writing" (EW, p. 19).

But why exactly do these methods work? What is it they do that reduces readers' well documented tendency to disagree, sometimes widely, in their assessments of essays? What, in the readers' reading, do these methods alter in order to ensure inter-rater consensus?

The obvious answer to these questions would seem, at first, to be that these methods reduce variation in the value systems being applied by readers to texts, either by limiting the group of readers to those who are likely to share value systems to begin with (as in methods one and two on my list), or by encouraging readers toward conformity by making them state explicitly the qualities in texts they will reward with high scores or penalize with low ones. This "obvious answer" undoubtedly does explain part of what the methods used to promote inter-rater reliability do. This answer, however, rests on an assumption, which also seems "obviously true" at first, that the readers exercise their various value systems on the same set of stable things, the texts--that when two readers read a text, that is,

they will first perceive it as "having" the same "qualities," as "saying" the same "things" in the same "ways"--and then they will decide, by applying their individual value systems, whether in their judgment the test says those things well or ill, whether the qualities it has are assets or liabilities.

Enough has been written in literary criticism of late concerning the role of readers as creators of literary texts, however, to cast doubt on the "obviousness" of the truth of these assumptions. Today I want to examine the work of two literary theorists, E. D. Hirsch and Stanley Fish, who have written about the activity of interpreting texts and specifically about the question of whether different readers share perception of the same thing when they look at a text. These figures represent two positions within critical theory that I believe are relevant to the issue of rater reliability in essay scoring: critical monism (which, in the sense I'm using it here, will refer to the critical position asserting that a text can be said to have only one correct meaning), and critical pluralism (which, in the sense I'm using it here, will refer to the position asserting that a text can have more than one correct meaning). I'll examine Hirsch's position first and summarize briefly his stand on the questions of assessment and rater reliability, as he outlines it in The Philosophy of Composition. Then I'll summarize contrasting views presented by Stanley Fish on the subject of texts, and connect those, also, to the issue of inter-rater reliability.

Let me stress at the outset, however, that neither of these theorists--neither Hirsch the monist nor Fish the pluralist--supports the "obvious" claim that a text "contains" only one meaning for readers to "extract" or "decode" from it, or the claim that readers will necessarily perceive the "same text" as "saying" the same "things." Although Hirsch concludes that there is one meaning of a text that may justifiably be considered the "correct" one, neither he nor Fish asserts that only one meaning is construable from a single text.

Hirsch, for example, states that

> the nature of the text is to have no meaning except that which an interpreter wills into existence. We, not our texts, are the makers of the meanings we understand, a text being only an occasion for meaning, in itself an ambiguous form devoid of the consciousness where meaning abides. (AI, 76)

Because texts are "in themselves ambiguous," they allow themselves to be read different ways, to "say different things" to different readers. For this reason, among others, we cannot rely simply on public consensus about what a text "says" to determine its meaning. According to Hirsch, such an idea is "based upon a fundamental error of observation and logic" (VI, p. 13), for no such "public consensus" exists about what a text "says." As Hirsch puts it, "If a text means what it says, then it means nothing in particular. Its saying has no determinate existence but must be the saying of the author or a reader" (VI, p. 13). Hirsch gets around the problems opened up by a text's ambiguity by assigning one of the text's many possible meanings primacy: the meaning intended by the author. He argues that authors write in the belief that their meanings are accessible, and readers read in the belief that they may grasp an author's meaning. Although readers cannot know for certain that they have done so, they can in practice come to entertain a meaning identical with the one the author intended. Language users develop confidence in these possibilities from childhood. In Hirsch's words,

> The faith that speakers have in the possibility of communication has been built up in the very process of learning a language, particularly in those instances when the actions of the interpreter have confirmed to the author that he has been understood. These primitive confirmations are the foundation for our faith in far less primitive modes of communication. The inaccessibility of verbal meaning is a doctrine that experience suggests to be false, though neither experience nor argument can prove its falsity. But since the skeptical doctrine of inaccessibility is highly improbable, it should be rejected as a working assumption of interpretation. (VI, p. 18)

Hirsch finds hope for evaluating writing reliably, then, in readers' ability to identify a writer's aims or intentions and then to evaluate how effectively a

written text fulfills them. Hirsch carefully distinguishes this kind of evaluation, which he calls "intrinsic evaluation" and which he suggests might be done reliably by competent readers, from the kind of evaluation that would score writing on bases other than the writer's success in achieving his or her intentions, and would even allow for scoring the intentions themselves. This latter kind of evaluation, which Hirsch calls "extrinsic evaluation," introduces differences in readers' "temperament and taste" (PC, p. 183) into the scoring process and thus prevents consensus among evaluators. Hirsch argues that these two legitimate modes of evaluation, which he traces to differences between Platonic and Aristotelian modes of judgment, have competed throughout the history of literary criticism; he claims that no set of "ideal proportions" for combining them into single holistic judgments of texts has ever been established, nor can one be.

Hirsch suggests, then, that "some of the variations in [readers' holistic] judgments, and in the weight attached to different aspects of writing, reflect different proportions of extrinsic and intrinsic judgments" included in the various readers' scores (PC, 183). Those proportions, Hirsch suggests, may be stabilized in the kind of training sessions that produce "a conformed group of readers," but such stabilization is arbitrary: "a different conformed group in a different time or place could not be relied on to yield similar results" (PC, 180). Since reliability of this sort is purchased at the cost of validity, it is an unsatisfactory solution to the problem of assessment, especially for research on writing. The solution Hirsch proposes is to isolate intrinsic evaluations from extrinsic evaluations in pedagogy, and to limit evaluation to the intrinsic sort in research. Hirsch proposes "relative readability" as a promising intrinsic standard by which to judge a writer's success. Assessments based on texts' relative readability could be confirmed by objective methods; they could then be considered "in a very legitimate sense, correct scores," according to Hirsch; and they could be used as a means of

certifying readers to evaluate student writing. These consequences would follow, that is, if readers could in fact identify an author's intentions reliably and use relative readability to measure how successfully the intentions were realized through the text.

Certain pluralist positions, however, question readers' ability to identify an author's intentions through a text. In the remainder of my talk, I want to examine one such radically pluralist position--that taken by Stanley Fish in some of his later work--in order to consider the questions it raises about the ideal of inter-rater reliability.

In a review of the Milton Variorum Commentary (IV, 1976), Fish states his opposition to several assumptions about texts, including "the assumption that there is a sense," and the assumption "that it is embedded or encoded in the text" (p. 473). Fish claims that these assumptions about texts undervalue the role readers play in constructing and reconstructing a text's meaning. Readers' activities, Fish claims, "are devalued because when they are thought of at all, they are thought of as the disposable machinery of extraction" (p. 474). Fish, like Hirsch, sees readers as creators rather than extractors. At first Fish argues for a kind of critical monism that would consist of outlining the "optimal reader's" responses to a given text--the "optimal reader" being that reader capable, by virtue of his or her "education, opinions, concerns and linguistic competences" of experiencing the text as the author intended it to be experienced. But a few pages later in the same paper, Fish dramatically undercuts any claim, including his own, that is based on the assumption that one can discover "in" a text its author's intentions: according to this "later Fish," perceptions of both formal units and intentions "are always a function of the interpretive model one brings to bear; they are not 'in' the text [.] . . . An intention, like a formal unit, is made when perceptual or interpretive closure is hazarded; it is verified by an interpretive

act, and . . . it is not verifiable in any other way" (p. 478). Thus one's sense of an author's intention is always traceable to one's interpretive strategy. And since, according to Fish, more than one legitimate interpretive strategy is open to readers, so is more than one sense of an author's intention.

In a section of his paper especially relevant to inter-rater reliability, Fish then confronts the question of why different readers ever perform similarly when reading the "same" texts. In answer to this question, Fish posits the notion of "interpretive communities":

> Interpretive communities are made up of those who share interpretive strategies not for reading (in the conventional sense [of decoding]) but for writing texts, [that is,] for constituting their properties and assigning their intentions. In other words these strategies exist prior to the act of reading and therefore determine the shape of what is read rather than, as is usually assumed, the other way around. (p. 483)

Fish argues that "interpretive strategies are not natural or universal, but learned"; one learns the rules of the various communities to which one is exposed. Fish thus can use the notion of interpretive communities to account not only for stability of interpretation among different readers, when it exists, but also for a single reader's ability to vary in his or her interpretation of a single text: in shifting interpretations, the reader simply shifts from one interpretive community to another. Since ways of interpreting can be "forgotten or supplanted, or complicated or dropped from favor" as well as acquired, the stability of interpretation provided by the existence of interpretive communities is temporary. It can be formed and dissolved, as the communities on which it is based form and dissolve in their turn. Fish sums up his position as follows: "Mine is not an argument for an infinitely plural or an open text, but for a text that is always set; and yet because it is set not for all places or all times but for wherever and however long a particular way of reading is in force, it is a text that can change" (NCLLDSA etc., 629). Texts are stable, to put it another way, only as long as

readers are. And thus, as a corollary, texts can be stabilized by stabilizing readers, by teaching them a common interpretive strategy, a common way of "seeing things in texts."

Fish's later position calls into question the ideal of inter-rater reliability, then, insofar as that ideal leads us to expect consensus across the boundaries of interpretive communities, for readers in different communities would be perceiving (or in Fish's term, "writing") different texts. By the same token, however, it allows for (though it does not ensure) the possibility of consensus within a community, and furthermore, it suggests a new way to account for readers' ability to move, through practice and discussion in a training session, from disagreement over texts to consensus about them. Such readers would be, among other things, establishing an interpretive community. The important issue is whether such an interpretive community's readings have validity beyond the community's boundaries, whether its judgments can be taken to represent the judgments of the set of competent readers as a whole--whether they may be said to be, in Hirsch's terms, "consistent with judgments about good writing in literate society at large" (PC, p. 177).

If we take Fish's position seriously, I believe we must deny that a single community's judgments have such validity. For unless we are willing to argue that competent readers, before their conformation in a training session, are not representative of "literate society at large," we must conclude that a degree of inconsistency is itself a consistent phenomenon in "judgments about good writing in literate society at large." A conformed community's readings are but one moment of stability in an essentially fluctuating process.

To accept such a position is not to argue that evaluations based on the readings of a single interpretive community are useless, however; one can imagine situations in which a reliable evaluation of a piece of writing by

readers drawn from a given interpretive community would be exactly what we and our students wanted and needed. Skilled readers are, for instance, exemplars of and even door-keepers to their respective interpretive groups. To the extent that cultural, professional, and academic groups and sub-groups are interpretive communities and we are conformed text "writers" within those communities, we have valuable information to offer students about the communities' ways of seeing: each community's evaluations become, in an important sense, its "school of thought." But such a community's judgment would not be the same thing as a valid general verdict on the quality of a piece of writing, and it could not fairly be represented as one.

In a radically pluralist world such as Fish's, grounds for encouraging interpretive agreement among readers in different interpretive communities do not exist: one can hardly argue that readers who have arrived at different correct readings of a text, and therefore in a sense have "read" different texts, should be encouraged to reach consensus on the quality of the different texts they've read. The possibility of such a consensus' occurring by accident, of course, cannot be ruled out; but neither can it be upheld as the theoretical ideal.

In interpreting these two positions for you today, I have behaved, as you may have noticed, as a monist. I've acted as though Hirsch's and Fish's texts have stable senses and it is possible for me to identify these writers' intended meanings on the basis of their texts and what I have perceived as being "in" them. So in one sense, then, I find myself at the end of my talk in the same camp as Hirsch, and in my actions I have implicitly denied Fish's claims validity. In doing so, I have enacted what is usually termed, reasonably, "common sense": most of us, most of the time, behave as though we inhabit such a monistic world and find doing so not only possible but advantageous.

Yet I find Fish's arguments, and the implications they have for the ideal of inter-rater reliability, troubling. For if things like Fish's "interpretive communities" account for any of the variance we find in untrained readers' scoring, then we are faced with an interesting problem. Unless we set ourselves the impossible task of identifying and polling the constantly fluctuating "interpretive communities" existing in the set of competent readers at any given time, we will not be able to distinguish variance arising from differences in readers' interpretive communities from random variance, though we would certainly wish to take account of the former and minimize the latter in any valid assessment of the effectiveness of a student's writing. Both kinds of variance would be reduced, however, when readers were trained in the ways listed earlier; both would flourish when readers were left untrained.

Hirsch and Fish offer accounts of interpretation that I believe are of interest to those of us concerned with the dynamics and politics of essay testing. Moreover, by turning to literary theory as well as to psychometrics for insights into the process of essay evaluation, we integrate aspects of our professional interests that are often, in practice, separated. Reading contemporary literary theory in terms of its relation to testing may lead us to conclude, however, that though the methods developed over the past century to promote inter-rater reliability seem to have brought the "reliability problem" closer to solution, they have also provided us with new problems in validity.

WORKS CITED

Cooper, Charles R. "Holistic Evaluation of Writing." Evaluating Writing: Describing, Measuring, Judging. Ed. Charles R. Cooper and Lee Odell. Urbana: NCTE, 1977.

Diederich, Paul B. Measuring Growth in English. Urbana: NCTE, 1974.

Fish, Stanley E. "Interpreting the Variorum." Critical Inquiry, 2 (Spring 1976), 465-485.

Fish, Stanley E. "Normal Circumstances, Literal Language, Direct Speech Acts, the Ordinary, the Everyday, the Obvious, What Goes without Saying, and Other Special Cases." Critical Inquiry, 4 (Summer 1978), 625-644.

Hirsch, E. D., Jr. The Aims of Interpretation. Chicago: Univ. of Chicago Press, 1976.

Hirsch, E. D., Jr. The Philosophy of Composition. Chicago: Univ. of Chicago Press, 1977.

Hirsch, E. D., Jr. Validity in Interpretation. New Haven: Yale Univ. Press, 1967.

Huddleston, Edith. "Measurement of Writing Ability at the College Entrance Level: Objective vs. Subjective Techniques." Journal of Experimental Education, 22 (1954), 165-213.

McColly, William. "What Does Educational Research Say about the Judging of Writing Ability?" The Journal of Educational Research, 64 (Dec. 1970), 148-156.

Stalnaker, J. "Question VI, the Essay." English Journal, College Edition, 26 (Feb. 1937), 133-140.

Stalnaker, J. and Stalnaker, R. C. "Reliable Reading of Essay Tests." School Review, 42 (1934), 599-605.

Stanley, Julian. "Reliability." Educational Measurement. 2nd ed. Ed. Robert L. Thorndike. Washington: American Council on Education, 1971.